[*Translating and the Computer 9*. Proceedings of a conference… 12-13 November 1987, ed. Catriona Picken (London: Aslib, 1988)]

Eurotra: past, present and future

Peter Lau

Commission of the European Communities, Luxembourg

Ten years ago Eurotra was just an idea discussed in an expert committee created by the Commission. After scanning the market in order to find a suitable MT system the members had realised that all available operational systems were American and based on a ten-year-old state-of-the-art in linguistics and computer science. In consequence, it seemed like a good idea to try to produce a European MT system based on up-to-date scientific and technological knowledge, and a few years later the Eurotra idea was clear enough to allow for a fairly detailed presentation of the nature and composition of the system. Such a presentation was given by a member of the expert committee, Margaret King, six years ago at the Aslib conference on Practical Experience of Machine Translation.

EUROTRA SIX YEARS AGO

The principal design criterion adopted from the very beginning was that of true multilinguality. Earlier systems took advantage of accidental similarities between the source and the target language, because they were entirely bilingual, and in this way they could avoid analysis of source language sentences (clauses, phrases) and synthesis of target language sentences (clauses, phrases) whenever these had identical structures. This meant that such systems were strictly bound to specific language pairs, and by the inclusion of new languages it was very difficult, if not impossible, to take advantage of the work already done.

In multilingual systems monolingual analysis and synthesis are separated from translation proper (called transfer) in order to allow for the inclusion of new

languages without disturbing or rewriting existing parts of the system and still taking advantage of the work already done on the other languages. The fundamental requirement of such a system is an extremely well-defined and stable interface between on one side analysis and synthesis and on the other side transfer.

The separation of analysis and synthesis from transfer leads to a modular system. In most transfer-based systems, however, the monolingual analysis and synthesis have been even further modularised, e.g. by separation of morphology and syntax and sometimes a separate semantic module. This is also the case in Eurotra, and the highly modular design goes very well with the decentralised nature of the project (in 1987 some 150 people dispersed over 20 sites in the Member States). On the other hand, modularity, especially when modules are developed independently at geographically distant sites, requires very detailed and minutely specific interfaces.

THE EUROTRA INTERFACE STRUCTURE (IS)

For obvious economic reasons the transfer component of a multilingual system should be kept as small as possible. This means in particular that structural transfer should be avoided and only lexical transfer allowed.

The best way of neutralising structural differences between languages seems to go through the creation of a fairly deep semantic representation of the source and target texts. In consequence, the Eurotra IS was defined from the outset as a semantic representation based on semantic roles of the kind found in *Fillmore's Case Grammar*. This representation should neutralise surface phenomena like passivisation, movements, extraposition, ellipsis and inversion and provide for a simple representation of pronouns, subordinate clauses, etc. If the structural differences between the source and target texts are neutralised, transfer is reduced to a simple translation of lexical units, and the deconstruction of the source text and construction of the target text are left to the monolingual analysis and synthesis components which work independently of the combination of languages in a concrete translation (i.e. one analysis/synthesis module per language covered by the system).

It was foreseen that the creation of a suitable semantic representation would require a lot of research work, and in the original planning of the project it was stipulated that the major bulk of research would concern the linguistic side, while the software side of the project should rather be based on already available products like high level programming languages (LISP, PROLOG), compiler compilers, etc.

It was also foreseen that linguistic research would have to stop at some point. Eurotra was not meant to solve all remaining problems in, e.g. semantics, especially not problems involving ‘knowledge of the world’ like some kinds of pronoun resolution. Consequently, Eurotra was never meant to produce fully

automatic high quality translation for direct use without post-editing. The goal was, and still is, to build a system which will produce translations of a reasonably good quality, and which is extensible, reparable and modular, hence allows for the inclusion of new languages and new and better representational theories on, e.g. morphology, syntax, semantics or translation.

THE REAL HISTORY OF EUROTRA

Thus, the nature and composition of the Eurotra system and the research and development (R&D) programme aimed at the construction of a prototype covering a limited number of subject fields (information technology) and text types (Commission documents) in all Community languages (20,000 lexical entries per language) were fairly clear six years ago. Nonetheless, it took another year until the Council of the European Communities adopted a Decision (4 November 1982) which provided the legal and budgetary basis for an R&D programme co-financed by the Commission and the Member States. The Council Decision foresaw a programme period of five and a half years subdivided in three phases:

1. The first (preparatory) phase in which the organisational infrastructure and the basis of the work on the national languages were to be established. The Member States were supposed to set up national groups to build analysis and synthesis modules for the official languages, and the Commission was supposed to provide linguistic and software specifications for this work.
2. The second phase in which a small system covering 2,500 lexical entries should be built and tested.
3. The third phase which should mainly be used for an expansion of the lexical coverage in order to reach 20,000 entries by mid-1988.

This schedule was not followed, however. By the end of the preparatory phase only four Member States had fulfilled their obligations according to the Council Decision, and the Commission still had not provided the eight temporary posts for the central Project Team which was supposed to produce specifications and coordinate the work of the national teams.

By the end of 1985 eight Member States had effectively joined the project, and the Commission had managed to provide the necessary specifications by means of contracts with researchers working in existing or prospective Eurotra groups. Internally, 1 January 1986 was considered the real starting date of phase 2, and by that date the Commission had already started the preparatory work in view of the expected inclusion of Spanish and Portuguese. It was decided that the extra workload on the national groups due to transfer from Spanish and Portuguese together with the delays already incurred made it necessary to ask for an extension of the programme period by eighteen months.

On 26 November 1986 Council adopted a decision which extended Eurotra to Spain and Portugal, increased the allocation of Commission staff from eight to fourteen and prolonged the programme till the end of 1989.

In the course of 1986 the remaining Member States joined the project (except that the Portuguese contract was not signed till March 1987), so it was not until 1987, eighteen months before it should have ended according to the original plan, that Eurotra became a full-scale project. Naturally, this had led to enormous differences in the progress of the various national groups, and we are still facing severe, if not insurmountable, problems in the process of synchronising their work.

Moreover, the Council Decision on the extension of the programme put a condition on the transition to the third phase: an external assessment panel should evaluate the project, and budget allocations for the third phase would depend on a positive outcome of this evaluation.

The assessment panel was set up during the first months of 1987, and its final report became available mid-October. It contains the following conclusions:

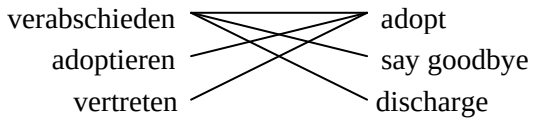
1. The Eurotra project model is the first of its kind, and it represents real progress in scientific collaboration within the European Community. Considering the difficult conditions under which the project has been carried out, it is surprising that it has reached full scale. The secondary but very important goals of dissemination of knowledge, transfer of know-how and training of computational linguists in the Member States have already been achieved.
2. The scientific approach is basically sound. It represents an attempt at exploiting the experience gained in other important MT projects in the world, and the project has made reasonable progress, particularly in view of the difficult conditions.
3. The management side has been less than optimally efficient. There has been a lack of qualified central staff and a tendency to let research dominate to the detriment of development.
4. The software is inefficient. Software development has been very much research-oriented and too little attention has been paid to the need for efficient software in the development of an MT system. In order to overcome this difficulty, industrial implementers should be involved in software development as soon as possible.
5. Recommendations:
   Eurotra should continue into the third phase. In this phase, however, development should be separated from research, and development should aim at a more immediate output, i.e. the construction of the prototype described in the first Council Decision and spin-off products. In order to achieve these aims, industrial partners should be involved as soon as possible.

THE PRESENT SITUATION

Today (November 1987) Eurotra is a full-scale project progressing towards the implementation of a first reduced prototype with 2,500 entries and a modest syntactic and semantic coverage.

Linguistic research within the project has reached a point where the problems are clearly delimited, and a well-defined and determined effort is being made to reach operational solutions. In the light of this development it now seems that the neutralisation of structural differences between the source and target texts in order to reduce transfer to a purely lexical level which was the main concern of the early theories on the Eurotra IS is the easy part of analysis and synthesis. The real difficulties are found in the need for disambiguation of the lexical units before, during and/or after transfer.

Let us take a look at the English verb ‘adopt’. A possible German translation would be ‘verabschieden’, but ‘adoptieren’ and ‘vertreten’ are also possible. On the other hand, ‘verabschieden’ may not only be translated into ‘adopt’, but also into ‘discharge’ and ‘say goodbye’. In both directions we have a one-to-many relation between source and target in translation:



A set of examples will illustrate the problems we meet when we try to decide on the correct translation in a specific context:

1. The council adopted the decision
   Der Rat verabschiedete den Beschluss
2. They adopted a child
   Sie adoptierten ein Kind
3. The council adopted an intransigent attitude
   Der Rat vertrat einen kompromislosen Standpunkt
4. Der Rat verabschiedete den Beschluss
   The council adopted the decision
5. Sie verabschiedeten sich von ihren Freunden
   They said goodbye to their friends
6. Der General wurde verabschiedet
   The general was discharged.

As far as possible, we try to employ syntactic criteria in disambiguation, because syntax is normally better understood and better described than semantics. Of the six examples mentioned, however, only example 5 may be uniquely characterised on syntactic grounds: if ‘verabschieden’ appears in its reflexive form (sich verabschieden) it must be translated as ‘say goodbye’. In the other five cases, we need to base our disambiguation on semantic criteria. The simplest

semantic criterion is founded on different readings of a word. It may be claimed, for example, that adopting a proposal is so different from adopting a child that our dictionary should contain two entries for 'adopt': adopt1 and adopt2, the latter specifying that it only combines with 'infantile' objects. Distinguishing readings from one another, however, only works for clearly polysemic words, and it is well known that polysemy is very difficult and partly non-operational as a working criterion, e.g. is there a marked difference between adopting a decision and adopting an attitude? Is this difference more or less marked than the one between adopting a decision and adopting a child?

In any case, it seems certain that the differentiation of various readings of a word will only solve some of our disambiguation problems. A large number will depend on semantic features subsumed under individual lexical entries. One example of this is 'verabschieden' in the sense of 'discharge'. In order to decide on the correct translation of 'verabschieden' we need to know whether the object is human or non-human, because only human objects may be discharged.

The most difficult problem of all lies in the distinction between, for example, adopting a decision and adopting an attitude. 'Decision' and 'attitude' may both be characterised as abstract, and it does not seem easy to find a distinguishing feature which will allow for a correct choice of the German translation in these contexts. We are experimenting with different interface structures and transfer strategies in order to solve such problems, and some viable proposals have emerged like leaving it to the synthesis module to choose between alternative translations on the basis of target language information (e.g. certain verbs only combine with certain classes of nouns which may have to be enumerated extensively), but we shall, no doubt, not solve all of them. Research in semantics for translational purposes will go on after Eurotra. For us the crucial point is that we should build a system which can incorporate new knowledge as it becomes available.

PERSPECTIVES FOR THE FUTURE

In the light of the recommendations arrived at by the assessment panel, optimal continuation of the project relies on successful exploitation of the development potential combined with further research in such a way that future results of the research work may provide input to the development process. The degree to which development should be separated from research within the present programme is still subject to negotiation. For the future, however, the Community Framework Programme for R&D 1987-1991 contains proposals for a Eurotra II, which covers industrial development towards a marketable product, and a Eurotra III, which is a research project of reduced scale compared to Eurotra I and aimed at the design of a more sophisticated semantic treatment based on artificial intelligence techniques, knowledge of the world, etc.